Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization

Maria Kadukova, Sergei Grudinin

This document is the edited Author's version of a Submitted Work that was subsequently accepted for publication in Journal of Computer-Aided Molecular Design after peer review. To access the published work see http://dx.doi.org/10.1007/s10822-017-0068-8.

Convex-PL – a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization

Maria Kadukova[1,2,3,4] and Sergei Grudinin*[1,2,3]

[1]Univ. Grenoble Alpes, LJK, F-38000 Grenoble, France
[2]CNRS, LJK, F-38000 Grenoble, France
[3]Inria, France
[4]Moscow Institute of Physics and Technology, Dolgoprudniy, Russia

We present a novel optimization approach to train a free-shape distance-dependent protein-ligand scoring function called Convex-PL. We do not impose any functional form of the scoring function. Instead, we decompose it into a polynomial basis and deduce the expansion coefficients from the structural knowledge base using a convex formulation of the optimization problem. Also, for the training set we do not generate false poses with molecular docking packages, but use constant RMSD rigid-body deformations of the ligands inside the binding pockets. This allows the obtained scoring function to be generally applicable to scoring of structural ensembles generated with different docking methods.
We assess the Convex-PL scoring function using data from D3R Grand Challenge 2 submissions and the docking test of the CASF 2013 study. We demonstrate that our results outperform the other 20 methods previously assessed in CASF 2013. The method is available at http://team.inria.fr/nano-d/software/Convex-PL/. **keywords**: Machine learning Molecular docking Protein-ligand interactions Scoring function Knowledge-based potential

Introduction

Development of new computational methods for the prediction of protein-ligand interactions is stimulated by the growing demand in drug discovery for the efficiency and accuracy of virtual screening, including molecular docking and binding affinity prediction applications. To computationally analyze numerous compounds, extremely fast techniques are required. These are typically used to preselect the most probable binders of a target protein for the further experimental investigation. Thus, improving accuracy and speed of such techniques is an active research field in structural bioinformatics.
The protein-ligand complex formation can be considered as a thermodynamic event described with the binding affinity constant, which is related to the binding free energy. Native and near-native binding poses should correspond the minima of the binding free energy. The aim of molecular docking methods is to somehow predict these poses with the corresponding energy values. The binding free energy for a protein-ligand complex is a combination of various terms including not only interactions between the protein and the ligand, but also solvation and entropic contributions. A rigorous computation of the binding free energy requires sampling of the complete configurational space, which is a very computationally demanding task due to its typically high dimensionality, which often makes

*sergei.grudinin@inria.fr

these approaches prohibitive. We should add that nowadays direct computation of the binding free energy of a molecular complex, or of the relative binding affinities associated with two ligands can be carried out with the thermodynamic integration technique. However, while this method can be applied to complexes consisting of few partners of small size, it generally reaches a prohibitively high computational cost, unless specific properties of the system can be exploited [1–3]. In response to this computational challenge, diverse approximation techniques that estimate the binding energy with a scoring function have been developed in the past years [4–13].

The currently existing scoring functions can be categorized into four groups based on the underlying principles of their work: physics-based methods, empirical scoring functions, knowledge-based potentials and descriptor-based scoring functions [14]. However, this classification is not rigorous and there are other scoring functions that combine several aforementioned concepts. *Physics-based* scoring functions are, perhaps, the most intuitively clear [15–19]. They are based on direct simulations of the possible physical effects of protein-ligand interactions. Despite a considerable progress in the force-field, quantum chemistry, and solvation models developments [20, 21], methods based on exhaustive sampling of the configurational space still require high computational resources and the physics-based scoring functions often suffer from unrealistic output energy values, which should be somehow scaled later on. *Empirical scoring* functions are a linear combination of several terms that represent energy contributions of possible interactions at the protein-ligand interface such as hydrogen bonds, hydrophobic effects, solvation, steric clashes and other terms that may vary from one scoring function to another [22–25]. These interactions are combined with weights that can be found by a multivariate regression analysis, which requires a training set with known binding affinity constants [26]. Therefore, empirical scoring functions strongly depend on the quality of the experimental data and may be biased. Nevertheless, they are widely used nowadays and often demonstrate good results on the test benchmarks [27]. For example, a very simple empirical scoring function was used in the popular open-source AutoDock Vina package [28], which was the basis for some recent scoring functions' development [29, 30]. We should specifically mention the random forests-based scoring functions [30], which recently became a popular method of choice. *Knowledge-based* potentials employ an assumption that statistical analysis of empirical (structural) data collected from protein-ligand complexes may uncover the differences between native and non-native binding poses [31–38]. Typically, these potentials are given as a sum of pairwise terms that are derived from the inverse Boltzmann statistical distributions of distances (or, generally, geometric features) between atoms of protein-ligand complexes [39, 40]. Knowledge-based potentials can also include other energy contributions, such as entropy and solvation. The training sets for these potentials contain only structural information, and are independent from the experimental binding affinity data. This allows to use larger training sets. Also, this approach avoids possible binding affinity ambiguities caused by experimental conditions [41]. Hence, knowledge-based potentials are thought to predict the binding poses rather than binding affinities, but can be successful in both of the tasks [11]. *Descriptor-based* scoring functions appeared in 2000th, being inspired by the growing popularity of machine learning techniques [42–44]. Although some authors refer to their descriptor-based methods as to an extension of empirical scoring functions, they can be definitely separated into a stand-alone category of scoring functions based on ideas coming from the quantitative structure-activity relationship (QSAR) techniques. Descriptor-based scoring functions rely on a set of various descriptors representing structural, topological, electro-statical, hydrophobic and other contributions to the protein-ligand interactions. One of the potential drawbacks of these methods is a lack of physical interpretation of the non-linear relations between the descriptors [45].

Generally, scoring functions are used for the following tasks – prediction of putative docking poses, relative affinity predictions, and the absolute affinity predictions [41, 46]. Most of the scoring functions show relatively good results in near-native docking poses prediction, however their estimated affinities are still low-correlated with the experi-

mental binding constants [8,13,35,47,48]. There are different benchmarks designed to assess the performance of scoring functions. The most notable ones are the Community Structure-Activity Resource (CSAR) benchmarks [9, 10, 49], benchmarks of its successor, the Drug Design Data Resource (D3R, `www.drugdesigndata.org`), and the Comparative Assessment of Scoring Functions (CASF) benchmark. All of them include tests for the docking poses prediction and the relative affinities prediction. Recently we have demonstrated performance of our scoring protocol in the CSAR, CAPRI and D3R exercises [50–53]. Here, we rigorously derive our scoring function for protein-ligand interactions, called Convex-PL, and assess its performance using data from D3R Grand Challenge 2 submissions and the docking test provided in the CASF 2013 study [13, 27]. We also compare it with other prediction methods. More specifically, we focus on the assessment of the ability of our knowledge-based function deduced solely from the structural data to predict the correct docking poses.

Compared scoring functions

Here we will briefly list the scoring functions that were assessed by the authors of the CASF 2013 study. For more details we refer the reader to the corresponding papers cited in this section or to the descriptions provided in Supporting Information of the CASF 2013 paper [13] and references therein. We should note that the vast majority of the scoring functions assessed in the CASF 2013 study are parts of proprietary chemical software packages. From now on, we will refer to them as to *function_name@package_name*.

The simplest scoring function assessed in CASF 2013 is ΔSAS. It scores protein-ligand complexes according to the difference of the solvent-accessible surface area (SASA) upon formation of the complex. Despite a deceptive simplicity of such a scoring with the ΔSAS function, it demonstrates a very good correlation with experimental results in the relative scoring tests, which will be shown below.

Several empirical scoring functions assessed in CASF 2013 are based on the ChemScore function proposed by Eldridge et al. [23] estimating the binding free energy with a combination of hydrophobic and metal interactions distance-dependent terms, hydrogen bonding distance- and angle-dependent terms and a term accounting for the flexibility penalties of frozen rotatable bonds. Glidescore@GLIDE [25] extends this approach with taking into account the charge of atoms involved in the hydrogen bonds formation. It also includes Coulomb and van der Waals' energy contributions, as well as terms describing hydrophilic-hydrophobic interactions and solvation effects. ChemScore@GOLD scoring function includes a term that measures internal ligand energy and a term designed to penalize steric clashes. In ChemScore@SYBYL, a term that estimates conformational entropies is added, and the metal interactions term depends on both angles and distances.

GoldScore@GOLD is a force-field-based scoring function given as a sum of hydrogen bonding, van der Waals and torsion contributions. It also contains additional terms that were not enabled in the CASF 2013 assessment. ASP@GOLD [32] is a knowledge-based scoring function, in which pairwise statistical potentials are mixed with internal energy and clash terms taken from ChemScore@GOLD. ChemPLP@GOLD [26] is an empirical scoring function based on a piecewise linear potential (PLP) for attractive and repulsive contacts. It also includes hydrogen bonding and internal energy terms from ChemScore@GOLD.

The CASF 2013 paper also assessed five scoring functions implemented in Discovery Studio [54]. LigScore@DS [55] is an empirical scoring function available in two versions, the best of which combines a Lennard-Jones 9-6 potential and terms corresponding to attractive protein-ligand contacts and the ligand buried polar surface area. PLP@DS [56] is an empirical scoring function, consisting of pairwise potentials for different types of interactions including hydrogen bonding and steric interactions. Jain@DS [57] is an empirical scoring function taking into account hydrophobic, polar attractive, polar repulsive, and solvation contributions along with an entropic term. PMF@DS is a statistical scoring function derived from the potentials of mean force by Muegge et al. [31, 58, 59]. LUDI@DS [22, 60] is one more empirical scoring function consisting of a sum of distance- and

angle-dependent hydrogen bonding and ionic interactions terms, and also contributions that depend on hydrophobic buried surface area and rotatable bonds of the ligand molecule. Interestingly, one of several versions of the LUDI scoring functions presented in Discovery Studio also includes a term that estimates aromatic-aromatic interactions. In addition to ChemScore@SYBYL, three more scoring functions implemented in the SYBYL package were assessed by the authors of the CASF 2013 study. D-Score@SYBYL includes only a Lennard-Jones potential and electrostatic interactions. PMF-Score@SYBYL is a knowledge-based scoring function based on the potential of mean force [31, 58, 59]. G-Score@SYBYL is based on the GoldScore@GOLD and includes terms estimating hydrogen bonding energy, van der Waals energy of the complex and internal energy of the ligand.

As for the MOE package [61], four scoring functions were assessed in the CASF 2013 study. London-dG@MOE is a scoring function designed to predict the binding free energy that consists of terms describing the flexibility entropy of the ligand, geometrical imperfections of protein-ligand and metal-ligand interactions and desolvation energy approximated with the London dispersion forces between solute atoms and a continuum solvent [62]. Affinity-dG@MOE also estimates the free energy of binding with a linear sum of interactions of certain types. ASE@MOE [63] is a sum of Gaussian overlap functions between the ligand atoms and alpha spheres, with a parameter characterizing each alpha-sphere as occupying the space which is either accessible for a ligand or not. Alpha-HB@MOE consists of two terms, one of which is based on the alpha spheres, as in ASE@MOE, while the other describing the hydrogen bonding. Finally, X-Score is an empirical scoring function [24] that includes van der Waals' interactions, hydrogen bonding, hydrophobic and deformation effects, which was trained on the PDBBind database complexes.

Method

Model of interactions

Let us consider P native protein-ligand complexes C_{i0}, $i = [1, P]$. For each native configuration of the complex (which is generally a co-crystal structure found in structural databases), we generate D non-native configurations (decoys) by applying rigid transformations to the ligand and obtain C_{ij} decoys with $j \in [1, D]$, where the first index indicates the protein-ligand complex and the second index indicates the generated decoys. Thus, for each complex we have $D+1$ conformations, 1 native and D non-native. Our aim is to find a *scoring functional* E such that the following inequalities hold,

$$E(C_{i0}) < E(C_{ij}), \quad \forall i \in [1, P], \quad \forall j \in [1, D] \quad (1)$$

This is a difficult problem in such a general formulation. In order to solve it, we need to make some simplifications. Thus, we represent the protein-ligand complex as a set of atoms, which are split into a finite number of types. We choose these types according to atoms' properties such as chemical element, aromaticity, hybridization state, and polarity. This results in a total of $M_1 \times M_2$ pairs of different interactions, with M_1 being the total number of protein atom types, and M_2 – the total number of ligand atom types. Then, we assume that E depends only on the distribution of the distances between the atoms, with one atom located on the protein and the other on the ligand. We also assume these interactions to be short-ranged, which can be neglected if the distance between two interaction atoms is larger than a certain cutoff distance r_{max}. This allows us to restrict the information extracted from the complexes to their interfaces. We use a cut-off distance value of 10 Å, which has been widely used in previous approaches [64–69], and which gave good results in our earlier experiments [50]. Finally, we assume that E is a linear functional of the following form,

$$E(n(r)) = \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} \int_0^{r_{max}} n^{kl}(r) f^{kl}(r)\, dr, \quad (2)$$

where $n^{kl}(r)$ are the *number densities of atom-atom pairs* at a distance r with the first atom of type k

located on the protein, and the second atom of type l located on the ligand, and $f^{kl}(r)$ are the unknown *interaction potentials* between the atoms of types k and l. In our method, we use the following functional form for the number densities $n^{kl}(r)$,

$$n^{kl}(r) = \sum_{ij} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(r-r_{ij})^2}{2\sigma^2}}, \tag{3}$$

where each distance distribution is represented with a Gaussian function centered at r_{ij} with the standard deviation σ of 0.4 Å. This value was fixed and adapted from our previous studies [37]. The sum is taken over all pairs of atoms i of type k and j of type l separated by the distance r_{ij} smaller than r_{max}, with atom i located on the protein molecule and atom i located on the ligand molecule. We should note that equation 2 is very similar to the standard widely-used scoring formula, where individual protein-ligand distance-dependent interactions are summed up. Indeed, we can re-write functional E in the canonical way,

$$E = \sum_{ij} \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} u^{kl}(r_{ij}), \tag{4}$$

with individual protein-ligand interactions $u^{kl}(r_{ij})$ given as a *convolution* of the interaction potentials $f^{kl}(r)$ with the Gaussians,

$$u^{kl}(r_{ij}) = \int_0^{r_{max}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(r-r_{ij})^2}{2\sigma^2}} f^{kl}(r)\ dr. \tag{5}$$

In order to determine the unknown potentials $f^{kl}(r)$, we decompose them along with the number densities $n^{kl}(r)$ in a *polynomial basis*,

$$\begin{aligned} f^{kl}(r) &= \sum_{q=0}^{\infty} w_q^{kl} \psi_q(r), \quad r \in [0; r_{max}] \\ n^{kl}(r) &= \sum_{q=0}^{\infty} x_q^{kl} \psi_q(r), \quad r \in [0; r_{max}], \end{aligned} \tag{6}$$

where $\psi_q(r)$ are the basis functions orthogonal on $[0; r_{max}]$, and w_q^{kl} and x_q^{kl} are the expansion coefficients of $f^{kl}(r)$ and $n^{kl}(r)$, respectively. The orthogonality of the basis function implies that the following identity holds,

$$\int_0^{r_{max}} \psi_i(r)\psi_j(r)\Omega(r)\ dr = \delta_{ij}, \quad r \in [0; r_{max}], \tag{7}$$

where $\Omega(x)$ is a non-negative weight function with the support on $[0, r_{max}]$, and δ_{ij} is the Kronecker delta function. Without loss of generality, we can assume that the basis functions are always scaled in such a way that the weight function is unity. Figure 1 shows two examples of basis functions orthogonal on $[0, r_{max}]$. Other basis functions, for example those orthogonal on $[0, \infty)$, can be used as well [70]. Thanks to the orthogonal basis functions, expansion coefficients w_q^{kl} and x_q^{kl} can be determined from the orthogonality condition (7) as

$$\begin{aligned} w_q^{kl} &= \int_0^{r_{max}} f_{kl}(r)\psi_q(r)\ dr \\ x_q^{kl} &= \int_0^{r_{max}} n_{kl}(r)\psi_q(r)\ dr \end{aligned} \tag{8}$$

Using expansions (6), the functional E can be rewritten as

$$E(n(r)) = \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} \sum_{pq}^{\infty} w_q^{kl} x_p^{kl} \int_0^{r_{max}} \psi_q(r)\psi_p(r)\ dr \tag{9}$$

Finally, to have a compact representation, and thanks to the orthogonality of the basis functions, the scoring functional E can be truncated up to the order Q as

$$E(n(r)) \approx \sum_{k=1}^{M_1} \sum_{l=1}^{M_2} \sum_{q=0}^{Q} w_q^{kl} x_q^{kl} = (\mathbf{w} \cdot \mathbf{x}), \quad \mathbf{w}, \mathbf{x} \in \mathbb{R}^{Q \times M_1 \times M_2} \tag{10}$$

We will refer to the vector $\mathbf{w}$ as to the *scoring vector*, whose value is to be determined, and to the vector $\mathbf{x}$ as to the *structure vector* that is computed from the structural data. We should note that the vector $\mathbf{w}$ defines the interatomic potentials $u^{kl}(r)$ for protein-ligand interactions. To conclude, equations 8 provide a projection from a 3D structure into the *scoring space* on $\mathbb{R}^{Q \times M_1 \times M_2}$, while equation 10 defines the scoring functional in this space.

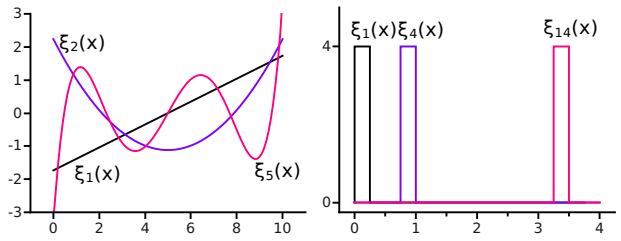

Figure 1: Basis functions orthogonal on $[0, 10]$. Left : Scaled Legendre functions of orders 1, 2, and 5. Right : Rectangular functions.

Geometric interpretation

Using the expansion of the scoring functional E provided by equation 10, we can reformulate the scoring problem 1: given P native structure vectors $\mathbf{x}_i^{nat}$ and $P \times D$ nonnative structure vectors $\mathbf{x}_{ij}^{nonnat}$, find such a scoring vector $\mathbf{w} \in \mathbb{R}^{Q \times M_1 \times M_2}$ that

$$\forall i = 1...P, \quad \forall j = 1...D \quad (\mathbf{x}_i^{nat} \cdot \mathbf{w}) < (\mathbf{x}_{ij}^{nonnat} \cdot \mathbf{w}), \quad (11)$$

or, equivalently,

$$\forall i = 1...P, \quad \forall j = 1...D \quad ([\mathbf{x}_{ij}^{nonnat} - \mathbf{x}_i^{nat}] \cdot \mathbf{w}) > 0, \quad (12)$$

which is a set of $P \times D$ *half-space equations* in $\mathbb{R}^{Q \times M_1 \times M_2}$ with P parallel separation hyperplanes defined by the common normal $\mathbf{w}$. Figure 2.3 schematically shows three groups of structure vectors separated by three parallel hyperplanes with a common normal $\mathbf{w}$.

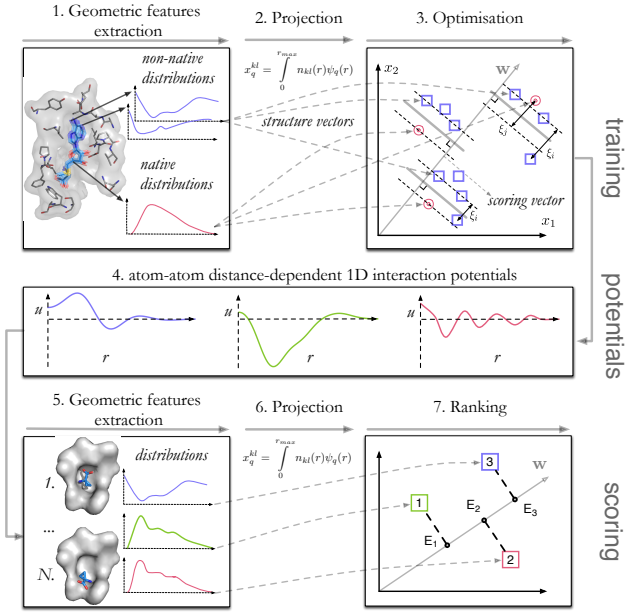

Figure 2: Schematic representation of the training stage (top) and the scoring stage (bottom) of the method.

The set of inequalities (12) can have zero, one or infinite number of solutions [71]. Generally, this is an ill-posed problem. To obtain a single solution, we rewrite it as a *soft-margin quadratic optimization problem* [72] with an additional quadratic *regularization* term,

$$\begin{aligned} &\text{Minimize (in } \mathbf{w}, b_i, \xi_{ij}\text{): } \tfrac{1}{2}\mathbf{w} \cdot \mathbf{w} + \textstyle\sum_{ij} C_{ij}\xi_{ij} \\ &\text{Subject to:} \\ &y_{ij}\left[\mathbf{w} \cdot \mathbf{x}_{ij} + b_i\right] - 1 + \xi_{ij} \geq 0, \;\; i = 1..P, \;\; j = 0..D \\ &\xi_{ij} \geq 0 \end{aligned} \quad (13)$$

Here, structure vectors $\mathbf{x}_{ij}$ are the same as in the above inequalities (11)-(12), index i runs over different protein-ligand complexes and index j runs over different conformations of the i-th protein-ligand complex. Particularly, protein conformations with $j = 0$ are native with the corresponding classifier $y_{i0} = +1$ and protein conformations with $j = 1..D$ are the decoys with the corresponding classifier $y_{ij} = -1$. Parameters C_{ij} can be regarded as regularization parameters, which control the importance of different structure vectors. To reduce the amount of adjustable parameters C_{ij} to a single regularization parameter C, we set values of C_{i0} for the native structure vectors to C, and the values of $C_{i1..D}$ for the non-native structure vectors to C/D. Then, we found the optimal value of the C parameter using the holdout cross-validation procedure [73]. The scoring vector $\mathbf{w}$, the offset vector $\mathbf{b}$ and the slack variables ξ_{ij} are the parameters to be optimized. Figure 2 shows a schematic workflow of the training and scoring stages of the presented method. We should note that the formulation 13 is, of course, not unique. We have also tried other regularization terms and other loss functions, but finally chose the quadratic regularization and the hinge-loss misclassification penalties because of many recent developments in the field of support vector machines [71,74], as we explain below.

Optimization algorithm

Solutions and properties of quadratic optimization problems similar to the one stated above (13)

have been extensively studied in theory of convex quadratic programming (QP) [71,74]. Using the notion of *Lagrangian*, the optimization problem (13) can be converted into its dual form [71,74], which is a convex QP problem with the objective function solely depending on a set of *Lagrange multipliers* λ_{ij},

$$\begin{aligned} &\text{Maximize (in } \lambda_{ij}\text{):} \\ &\mathcal{L}(\lambda_{ij}) = \sum_{ij} \lambda_{ij} - \frac{1}{2} \sum_{ij} \sum_{kl} y_{ij} y_{kl} \lambda_{ij} \lambda_{kl} \mathbf{x}_{ij} \cdot \mathbf{x}_{kl} \\ &\text{Subject to:} \quad \begin{array}{c} 0 \le \lambda_{ij} \le C_{ij} \\ \sum_j y_{ij} \lambda_{ij} = 0, \quad \forall i \end{array} \end{aligned} \tag{14}$$

We should note that this dual QP problem is convex because the corresponding matrix $Q_{(ij),(kl)} = y_{ij} y_{kl} \lambda_{ij} \lambda_{kl}$ is positive semi-definite. Thus, to solve it we can apply efficient techniques developed in the theory of convex quadratic optimization. In particular, the dual representation (14) of the original primal QP problem (13) allows us to break the original QP problem into a series of smaller subproblems. More precisely, various decomposition techniques have been developed to reduce requirements of QP solvers on the size of available RAM [75–78]. Here, we employ a *block-decomposition technique* and analytically iteratively maximize the Lagrangian with respect to pairs of multipliers according to *sequential minimal optimization* (SMO) algorithm [77]. We should specifically emphasize that the initial values of our potentials were set to zero and no inverse Boltzmann statistics was used during the optimization.

Vectors $\mathbf{x}_{ij}$ for which $\lambda_{ij} > 0$ are called *support vectors*. Once the dual QP problem (14) is solved and the optimal Lagrange multipliers λ_{ij} are found, we can express the optimal solution of the original primal problem (13) (the scoring vector) as a linear combination of the support vectors,

$$\mathbf{w} = \sum_{\text{support vectors}} y_{ij} \lambda_{ij} \mathbf{x}_{ij} \tag{15}$$

Atom types

Convex-PL describes the ligand and protein atoms with 41 and 23 types, respectively. To make the typization of a ligand, we use our recently developed Knodle (KNowledge-Driven Ligand Extractor) library [79]. It allows a conversion of a ligand given in the PDB format into either the Tripos Mol2 format or to an extended format of 164 types, based on the fconv extended type set that aim to represent chemical properties of different atoms [80].

To reduce the dimensionality of the chemical space, we selected several sets ranging from 30 to 52 atom types out of the initial 164 types using the similarity of the corresponding radial distribution functions. The initial version of our potential contained 52 ligand atom types. In our previous computational experiments [50, 52] we used the version with 48 types. However, we later realized that these numbers are too large for the current training set, as the potentials for some specific types contained oscillations at large interaction distances, which was a reason to merge them with more frequently occurring types. Finally, we chose the typization with 41 atom types because this provided the best cross-validation success rates on the control set. These are 8 carbon types, 14 nitrogen types, 7 oxygen types, 3 sulphur types, 2 phosphorus types, and 7 types describing halogens. These types are listed in Table S1 of Supporting Information.

As proteins contain a reduced chemical subspace compared to small molecules, we used for them a smaller typization set consisting of 23 reduced types. We should note that we did not include explicit hydrogens in our atom types in order to reduce the total number of atom type combinations and also to avoid possible errors in their assignment. Indeed, hydrogens are rarely resolved experimentally and typically included into the structures a posteriori. We should also add that our model does not contain directional (or angle-dependent) terms, that would significantly increase the size of our feature-space.

Figure 3 shows the matrix of numbers of pairwise contacts between these 23 protein and 41 ligand atom types computed for the training set. As it can be expected, protein types that correspond to the protein backbone and carbon atoms are very frequent. However, we definitely lack statistical data for the selene atoms that occur in modified protein residues. As for the ligand atom types, the rarest are those that correspond to ionic halogens, one of

the phosphorus types and one of the nitrogen types. Generally, a small number of available interactions for a certain pair of atom types in a training set may result in unnatural shapes of the obtained potentials. A possible way to overcome this problem will be to modify the regularization coefficients in the optimization problem (13), such that the contribution of a particular pair of types to the loss function becomes proportional to its frequency in the training set. However, we did not study this possibility.

Training and control sets

To train our model of binding free energy given by Eq. 10, we collected structural information from the PDBBind [81, 82] database, which provides experimentally determined protein-ligand complexes deposited in the Protein Data Bank supplied with the measured binding affinity data. We should note that we did not use the binding information for our analysis. Overall, PDBBind contains three-dimensional structures of resolution equal to or better than 2.5 Å of complexes found in Protein Data Bank (PDB) along with the corresponding binding data, which includes K_d, K_i, and IC_{50} values. To construct the PDBBind database, its authors manually examined the primary references for each protein-complex and collected experimentally determined binding affinity data. These constitute the "general set" of the database consisting in total of 14,620 complexes (as of release 2015), including protein-ligand (11,987), nucleic acid-ligand (109), protein-nucleic acid (717), and protein-protein complexes (1,807). Then the authors of PDBBind additionally compiled the "core set" to select 195 protein-ligand complexes as a high-quality benchmark for evaluating various docking/scoring methods. To do so, they applied a number of filters to the "general set" regarding binding data, crystal structures, as well as the nature of the complexes systematic, and did a systematic, non-redundant sampling of the obtained results. As for us, to construct the training set, we used randomly chosen 80% of protein-ligand complexes from the "general set" of PDBBind release 2015, excluding 195 complexes that intersect with the "core set" of the same database, as the "core set" forms the CASF 2013 benchmark. This resulted in 9,372 structures in the training set. The remaining 20% of protein-ligand complexes from the "general set" of the PDBBind release 2015, excluding 195 complexes from CASF 2013, formed the "control" set. We used this set to only adjust free parameters in our prediction model during its training. Both the control and training sets included decoys generated by the same procedure, which is described below.

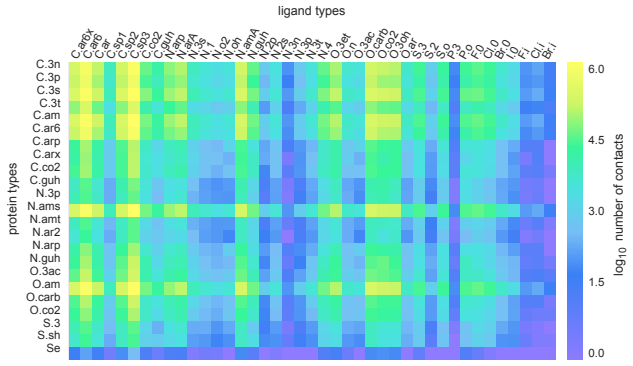

Figure 3: Numbers of pairwise contacts between the protein and ligand atom types, as found in the training set. All numbers are shown in a log-10 scale.

Generation of decoys

Generation of the decoy conformations is based on the idea from our recent work [37], where we demonstrated that structural information collected from native and *near-native* protein-protein complexes allows to construct powerful predictive models of protein-protein interactions. The *near-native* complex was defined as a complex with a small ligand-RMSD value, typically about 1 Å, with respect to the native structure. For the protein-protein case, we previously generated *near-native* protein-protein conformations using deformations of the native structure along some finite number of collective motions computed using the Normal Mode Analysis [37]. For the protein-ligand case, however, in order to generate *near-native* conformations, we consider ligand molecules as rigid bodies and rotate them about some axes defined such that the ligand-RMSD from the native pose is kept constant. To do so, we chose six axes inside a unit sphere corresponding to its icosahedral tessellation. More pre-

cisely, to generate the axes, we first aligned the principal axes of inertia of the icosahedron with the coordinate axes, and then connected its opposite vertices as it is shown in Fig. 4A. The weighted RMSD for a pure rotation about axis **n** by an angle α is [83]

$$\text{RMSD}^2 = \frac{4}{M}\sin^2\frac{\alpha}{2}I(\mathbf{n}), \tag{16}$$

where the ligand molecule of mass M is considered as a rigid body, whose inertia tensor relative to the axis **n** passing through its center of mass is given as

$$I(\mathbf{n}) = \mathbf{n}^T I \mathbf{n}. \tag{17}$$

To obtain a set of decoys with a certain ligand-RMSD from the native structure, we first rotated the ligand about each rotational axis by an angle of $\pm\alpha$ and then translated along coordinate axes by the lengths of $\pm$RMSD. Thus, for each native structure we generated 18 decoy conformations, which means that the total amount of the training structure vectors was $(18+1)\times 9,372 = 178,068$. Figure 4B shows an example of generated decoys for RMSD of 0.5 Å.

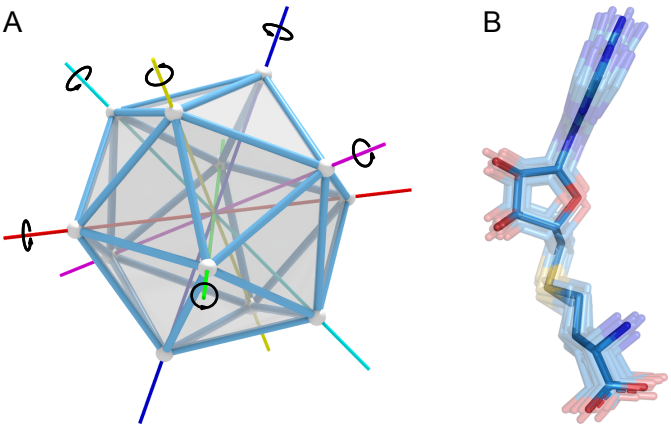

Figure 4: Decoys generation procedure. A : Six icosahedral axes about which we rotate the ligand. B : An example of a native ligand configuration with the corresponding 18 decoys generated with RMSD of 0.5 Å. These are 12 rotational decoys and 6 translational decoys.

In order to determine the optimal value of decoy RMSD, we carried out the two-fold cross-validation procedure using the training and the control sets. More precisely, we solved optimization problem 13 using structure vectors from the training set and measured the accuracy of the predictions on the control set. We exhaustively repeated this procedure while scanning through different parameters of RMSD and the regularization parameter C. Figure 5 shows the accuracy of our model with respect to the two adjustable parameters. We can see that, in principle, we can use any value of RMSD for decoy generation inside the [0.2 Å, 1.0 Å] interval, provided that the value of the regularization parameter C is chosen accordingly. Thus, for all further experiments we chose the value of RMSD equal to 0.5 Å. We want to emphasize that we specifically did not generate decoys with molecular docking programs. Indeed, our goal is to obtain a scoring potential that is unbiased with respect to methods for the docking pose generation. Thus, we chose the constant-RMSD rigid deformations for the decoys.

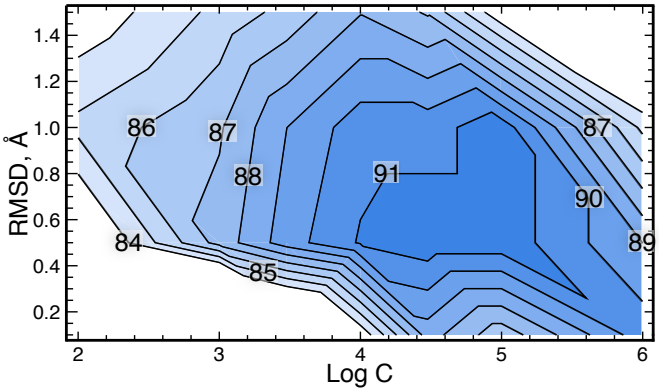

Figure 5: Success rate of the scoring function (in %) on the test set as a function of two adjustable parameters, RMSD for decoy generation, and the regularization parameter C. The success rate was computed as the percentage of correctly predicted decoys generated at RMSD=2 Å.

Repulsion term

We should note that the native complexes in the training set and also the generated decoys did not have atom-atom statistics at short interatomic distances. Consequently, the optimization procedure, or more precisely, its regularization term, left zero values of the scoring vector $\mathbf{w}$ and potentials $u^{kl}(r)$

at these distances, typically within two or three angstroms. To use our scoring function with the structures that contain atomic clashes, we manually filled these regions of $u^{kl}(r)$ potentials with artificial barriers of $\nu^{kl}r^{-2}$ shape. We adjusted the fitting coefficients ν^{kl} for each potential $u^{kl}(r)$ to match the first maximum of the $u^{kl}(r)$ curves. Also, to represent a soft repulsion at a zero separation distance, we replaced the barriers $\nu^{kl}r^{-2}$ at distances [0 Å, 0.4 Å] with a linear function, such that their values and first derivatives match at a distance of 0.4 Å.

CASF Benchmark

In this work, we assessed the obtained scoring function using the CASF 2013 benchmark examples. Overall, the CASF benchmark consists of 195 complexes formed by 65 proteins. For each of the proteins, three complexes with three different ligands of weak, medium and strong binding affinities are given. The preparation of this test set, along with a discussion about its advantages and disadvantages is reported by Li et al. in the CASF 2013 paper [27]. Four tests were suggested by the authors of CASF 2013 to assess the abilities of scoring functions to meet requirements of current pharmaceutical tasks that arise in academia and industry. However, we have only assessed our scoring function using the "docking" test. In this test the aim is to predict the best near-native docking poses. The CASF 2013 benchmark also presents the docking test results for 20 popular scoring functions that we have described above. It also repeats the test on three subsets of the test set, corresponding to three descriptors. These are the number of rotatable bonds in the ligand molecule, the fraction of ligand solvent-accessible surface area buried upon binding, and a particular descriptor representing the hydrophobic and hydrophilic properties of the protein binding pocket.

D3R Grand Challenge 2 Benchmark

We have also taken an opportunity to assess the pose prediction power of Convex-PL on the recently published user submissions to the pose prediction part of Stage 1 of the Drug Design Data Resource (D3R, `www.drugdesigndata.org`) Grand Challenge 2 [53,84]. This resource provides unpublished experimental data for testing protein-ligand docking algorithms and protocols. The pose prediction part of the Stage 1 of the Grand Challenge 2 was focused on blind docking of 36 ligands of different chemical series (benzimidazoles, isoxazoles, spiranes, sulfonamide and unclassified) to the farnesoid X receptor (FXR) target. High-quality experimental data was provided by Roche. Participants were allowed to submit no more than five predictions for each of the 36 ligands. After the Challenge was finished, all the user submissions including both the receptor and the ligand structures with corresponding RMSD values to the co-crystal structures became publicly available. For each of the target the Challenge participants submitted 169 – 184 of docking predictions. One of the ligands was excluded from the evaluation due to crystallographic artifacts, resulting in a set of protein-ligand complexes for 35 ligands that we used for re-scoring.

Computational details

We used the following parameters of our method, expansion order $Q = 25$, rectangular basis functions (see Fig. 1), number of protein atom types $M_1 = 23$, and number of ligand atom types $M_2 = 41$. Thus, the dimensionality of the feature space is $Q \times M_1 \times M_2 = 23,575$. We should note that we fixed the expansion order Q to r_{max}/σ =10 Å/0.4 Å. The size of the training set was $P \times (D + 1) = 9,372 \times 19 = 178,068$, which is significantly larger than the dimensionality of the feature space. However, some interatomic interactions for rarely occurring atom types still require more data. All the code was written using the C++ 11 language statically linked with the Knodle library [79]. We ran the tests on a Linux machine with 16Gb DDR3 RAM and Intel(R) Xeon(R) CPU E5-2609@2.40GHz. We also provide an executable for the Mac OS operating system.

Results and discussion

Obtained potentials

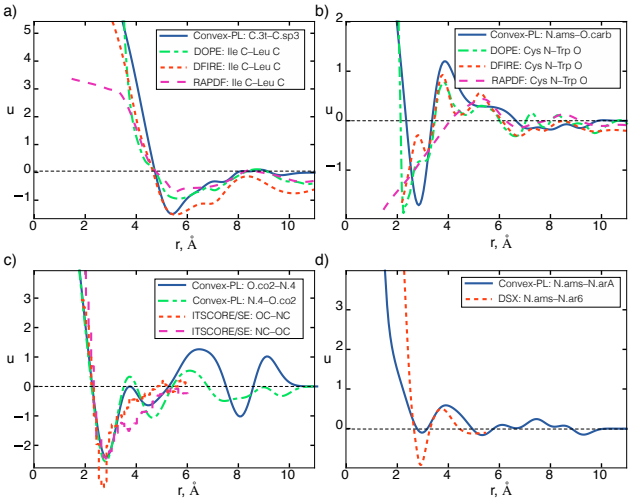

Figure 6: Comparison of the Convex-PL potentials with DOPE, DFIRE and RAPDF potentials for protein-protein interactions (a-b), with ITSCORE/SE (c) and with DSX (d) potentials for protein-ligand interactions. For the protein-ligand potentials, the two atom types in a pair correspond to the protein and ligand atoms, respectively. All the reference potentials' values were adapted from the plots found in the literature. The following interactions are plotted, a) sp3 carbon with sp3 carbon; b) secondary amide nitrogen with amide oxygen; c) negatively charged oxygen with positively charged nitrogen and vice versa; d) secondary amide nitrogen with aromatic nitrogen.

Starting from the initial vectors filled with zeros, after the optimization we obtained the scoring vector $\mathbf{w}$, which we also converted into 23×41 interatomic potentials $u^{kl}(r)$ for the sake of comparison with other methods. Some of these are shown in Fig. 6. In Figure 6 we plot our $u^{kl}(r)$ potentials together with several other potentials for relatively similar protein-protein and protein-ligand atom types that we found in the literature for DOPE [85], DFIRE [86], RAPDF [87], ITSCORE/SE [40] and DSX [36] scoring functions. Overall, it can be seen from this figure that Convex-PL predicts the first minimum and maximum peaks at similar locations compared to the other potentials. The difference between the first energy minimum locations in Fig. 6B may be caused by the fact that for protein-ligand interactions these interatomic separation distances are larger compared to similar protein-protein interactions. In Figure 6C we can also see two peaks for the Convex-PL potential with a protonated nitrogen on the ligand molecule that look unphysical. One of the possible explanations for this behavior is the difficulty of the correct type assignment for this type (N.4) on the ligand atoms. More precisely, this atom type can be easily mixed up with a sp3 nitrogen type, as the precise type assignment requires the presence of hydrogen atoms in the structure, while the corresponding lysine nitrogen on the protein molecule is considered to be protonated by default.

CASF 2013 docking test

For this test, 195 sets of decoys were provided by CASF 2013 for each protein-ligand complex. Each set contains 50-100 decoys with RMSD values smaller than 10 Å, generated by three different algorithms. The goal of the test is to predict the best near-native pose for each of the 195 complexes. Figure 7 shows the overall results of our Convex-PL potential with respect to some other top-performing scoring functions. Here, the success rates of finding top-1, 2 and 3 poses within RMSD values of 2 Å are shown. Results with the native and near-native structures excluded from the test set are listed in Table 1.

Figure 7 clearly demonstrates that Convex-PL shows the best performance in the detection of binding poses, with the success rates of 88.2%, 91.8% and 93.3% when predicting top-1, top-2 and top-3 poses, correspondingly, within RMSD of 2 Å. For example, it fails to identify top-3 poses in 9 out of 195 cases. However, two complexes out of these 9 should be considered individually, as for 3pxf and 3f3a complexes our potential tends to prefer decoys that are docked to the second binding site of the corresponding proteins. We should notice here that the authors of the CASF 2013 study visually verified the electron densities of the binding pockets if the structure factors were available. Notably, those complexes where the ligand electron density could be equally well fitted to multiple binding sites, were not included into the benchmark.

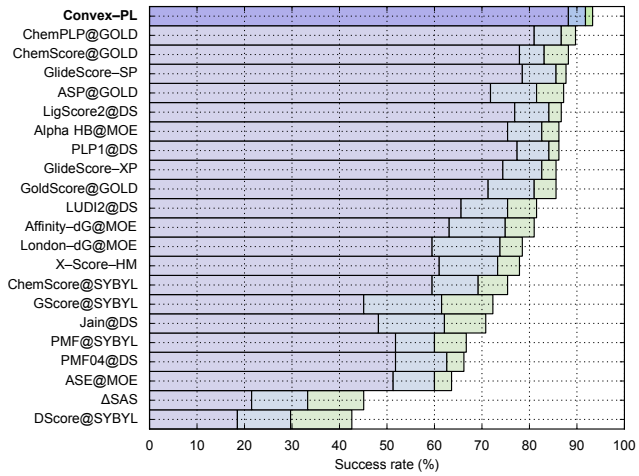

Figure 7: Success rates of finding top-1, -2 and -3 near-native structures within RMSD of 2 Å. Results are sorted by the top-3 success rates.

To have a more detailed picture of the performance of the Convex-PL potential in the docking test, we examined the scoring profiles, or "scoring funnels", of the 195 sets of docking decoys. These are the values of the Convex-PL potential as a function of RMSD of the corresponding docking poses. For a well performing scoring function we should expect a high correlation between the scores and RMSD values. Figure 8 shows several examples of these for both correctly and incorrectly predicted structures. In this figure, the top row demonstrates profiles for which the top-1 poses had RMSD smaller than 1 Å. Correspondingly, the bottom row demonstrates profiles for which none of these poses was in the top-1 scored configurations. We should note that in the two subsets we have proteins with several spatially proximal binding pockets containing the same ligands, such as the 3n86 and 3f3a complexes. As we can see in the plots, scores that belong to these alternative binding sites are separated into small clusters. A low correlation can be seen for the 3owj complex, whose ligand consists of several conjugated aromatic rings. This ligand is often predicted backwardly, which, nevertheless, seems to be not equivalent in terms of electron density because of asymmetrically located oxygen and nitrogen atoms. For 3ueu, whose ligand is an aliphatic chain consisting of sp3 carbons, only a 2 Å RMSD quality structure was successfully predicted. The rest of the complexes from the bottom row of Fig. 8 contain a large flexible molecule with several etheric rings (2qmj) and a sulphur-containing molecule with polar substituents and an etheric ring (3l4u), for which Convex-PL preferred a conformation with a rotated tail. Overall, in contrast to the scattered plots of configurations at large RMSD values, near-native scores exhibit a high correlation with RMSD values for all the plots in Fig. 8. We should mention that we should expect this high correlation because we trained our scoring function on decoys with small RMSD values.

In a practical scenario, due to conformational changes, modeling by homology or other inaccuracies in the receptor structure, it is often not possible to obtain near-native ligand poses with RMSD < 1 Å. This is clearly demonstrated in the recent D3R and CSAR challenges. Therefore, we have also evaluated the performance of Convex-PL on the CASF 2013 benchmark docking test with excluded native and near-native poses. These results are listed in Table 1. For example, if we exclude from the CASF test set both the native and the near-native poses within RMSD values of 1 Å, the top-1, top-2, and top-3 Convex-PL pose prediction success rates drop down to 57.4%, 70.1%, and 81.5%, respectively, for the poses found within 2 Å. These rates become 75.4%, 84.1%, and 91.3% for the poses found within 3 Å. We should note that these rates are not normalized to the total number of ligands with docking poses of a certain quality.

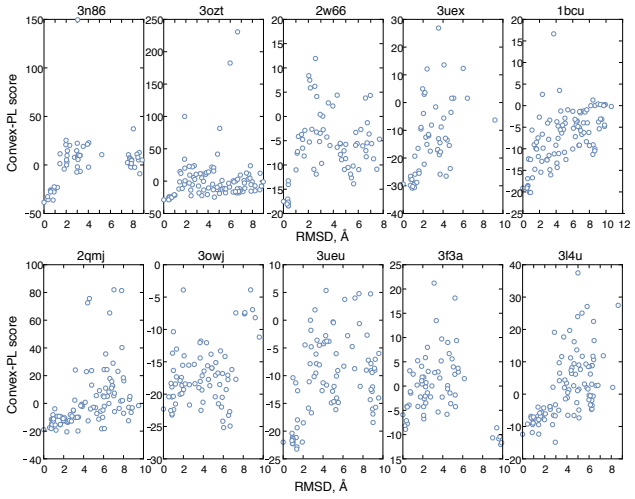

Figure 8: Convex-PL scores versus RMSDs. Top: Convex-PL score versus RMSD of a docking pose for correctly predicted structures (top-1 has RMSD < 1 Å). Bottom: Convex-PL score versus RMSD of a docking pose for incorrectly predicted structures (top-1 has RMSD > 1 Å).

The PDBBind "general set" contains a number of proteins homologous or even identical to those constituting the "core set", from which the CASF 2013 benchmark is constructed. To be confident that the presence of these proteins does not cause overfitting, we performed an accurate and computationally expensive leave-one-out cross-validation analysis. More precisely, for each of the 65 proteins from the "core set", we generated its own reduced dataset consisting of the "general set" without the protein's homologues. To do this, for each of the 65 proteins we detected its homologues in the "general set" using the 80% sequence identity criterion as computed by the BLASTP program of the BLAST+ package [88]. The list of the excluded proteins can be found in Table S2 of Supporting Information. After generating the 65 datasets, we divided them into the training and the control parts in the same manner as it was described above and ran 65 individual optimization processes. The resulting scoring functions were assessed on the docking test. All of these came to the same results as before, producing errors on a set of complexes that remains constant regardless the proteins excluded from the training set. Therefore, we can state that our scoring function is unbiased with respect to the proteins used in the CASF 2013 benchmark.

D3R Grand Challenge 2 docking test

Although Convex-PL demonstrated an excellent performance in the "docking" test of the CASF 2013 benchmark, we also decided to assess it on a more realistic docking problem. More precisely, for the second docking test we chose a diverse set of user predictions for the recent Stage 1 of D3R Grand Challenge 2. These contain complexes of the FXR apoprotein with 35 ligands of different chemical series. We re-scored all the user-submitted decoys for each of these 35 ligands, and obtained the results listed in Table 2. For each ligand we had 169 – 184 available docking poses with native complexes excluded from this test set. We normalized the success rates to the total number of ligands that had successful predictions of a certain quality. More precisely, D3R Challenge participants succeeded to predict poses within RMSD of 1 Å, 2 Å, and 3 Å for only 24, 33 and 34 out of 35 ligands, respectively.

We should note here that in such realistic docking problems the quality of predictions critically depends on the geometry of the receptor molecule. Indeed, dependent on the prediction quality and the number of the top-scored predictions, we are able to find near-native poses for all or almost all of the benzimidazole ligand targets. These were usually docked by the Challenge participants to a set of benzimidazole-containing receptors that are widely available in the Protein Data Bank. Among the top 5 poses with RMSD < 3 Å we were not able to find a solution for only four ligands, FXR_1, FXR_2, FXR_4, and FXR_23. These were among the most challenging targets for the Challenge participants with the mean RMSD values of the best submitted poses equal to 5.38 Å, 5.82 Å, 4.9 Å, and 6.15 Å, respectively. We should note that in this test we did not cluster or remove geometrically similar docking poses, as our aim was to assess the performance of Convex-PL. In a realistic docking exercise, however, a standard procedure would be to remove spatially proximate docking poses from the pool of top-5 docking solutions.

Evaluation on subsets

As we have mentioned above, the CASF 2013 benchmark provides subsets of the 195 complexes according to some of their chemical properties. Subsets A1–A3 correspond to smaller (A1), medium (A2) and higher (A3) number of rotatable bonds of a ligand. Unlike all the scoring functions, except GlideScore-SP, Convex-PL demonstrates almost the same docking power on the three subsets, as it is shown in Fig. 9.

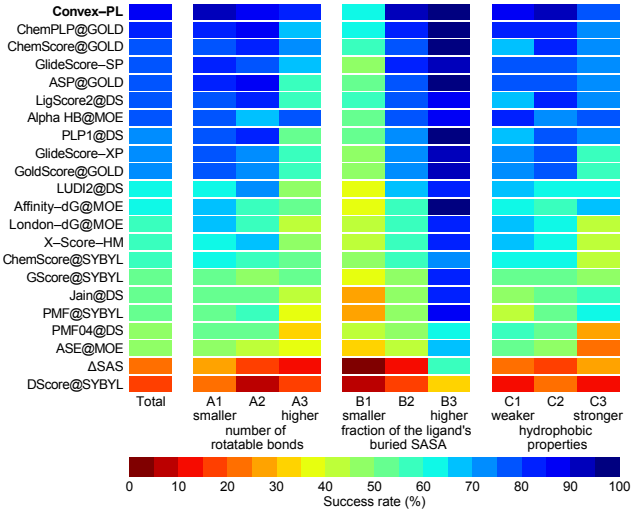

Figure 9: Pose prediction results of the top-1 structures with 2 Å RMSD quality on subsets. The leftmost column represents the success rate obtained on the whole set of 195 proteins.

Subsets B1–B3 represent smaller (B1), medium (B2) and higher (B3) fraction of the buried surface accessible area of a ligand. In general, the results on the B subsets show an inverse correlation with the results of the A subsets, as it is seen in Fig. 9. Indeed, most of the molecules with a small number of rotatable bonds are also small in size and thus deeply fit into the binding pockets with a larger fraction of buried solvent-accessible surface area. As for Convex-PL, it follows this trend, but the docking results on the B1 subset are considerably worse compared to the B2 and B3 subsets. One of the possible reasons of more frequent failures on the B1 subset is a smaller amount of protein-ligand interactions because of the larger exposed ligand surface.

The last three subsets, C1–C3, refer to more hydrophilic (C1), intermediate (C2), more hydrophobic (C3) properties. Here, Convex-PL shows almost the same results on the three subsets, with the best success rates produced on the intermediate C2 set, and slightly worse performance on the hydrophobic ligands. Overall, we can conclude that the performance of Convex-PL on the subsets has many similarities with the results demonstrated by other scoring functions, meaning that our method is not free from the common scoring functions difficulties.

Conclusions

In this paper we presented a machine-learning approach to train a free-shape distance-dependent protein-ligand scoring function. Distinct features of our approach are the following. First, we do not impose any functional form of the scoring function. Instead, we decompose it into a polynomial basis and deduce the expansion coefficients from the knowledge base. Second, for the training set we do not generate decoys with molecular docking packages, but use constant RMSD rigid-body deformations of the ligands inside the binding pockets. Therefore, the obtained scoring function is unbiased with respect to methods for the docking pose generation and can be generally applied to ensembles of molecular conformations generated with different docking methods. Third, for the optimization step, we use a quadratic programming formulation with the regularization term that aims to reduce possible overfitting to structural data. Our optimization problem is convex, and thus can be efficiently solved using multiple optimization techniques.

We have demonstrated the superior behavior of our potential in the docking test of the CASF 2013 benchmark examples, as compared to other 20 scoring functions assessed by the authors of the benchmark. We have also assessed our scoring function on a diverse set of user-submitted docking poses for the D3R Grand Challenge 2. Here, we obtained the success rates only slightly lower than the ones in the CASF 2013 benchmark despite the fact that all the co-crystal conformations of the receptors were unknown to the Challenge participants.

Several ways to improve our method can be envisaged. For example, low pose prediction results for the ligands with a small fraction of buried SASA suggests that including explicit interactions with solvent can help in this case. Second, we can reconsider the composition of the training set. Indeed, currently we generate the non-native configurations for each of the protein-ligand complex with different positions of the same ligand forming the complex. On the one hand, it allows us to construct a knowledge-based potential that is unbiased to different molecular docking packages, which may corrupt the solution of the machine learning problem

towards finding some decoy generation-related patterns. On the other hand, adding non-native ligand poses from the other complexes may help our scoring function to be applicable to screening tasks. Finally, we should add that including information about experimental binding affinities into the optimization problem will help to develop a scoring function specifically for screening applications. Indeed, ConvexPL is a pairwise-additive scoring function based solely on a large number of structural features extracted from protein-ligand interfaces. This allows a rather high flexibility when adjusting the weights of these features and may result in unrealistic relationships between the scores of different complexes. Extending the optimization problem with terms that penalize the divergence from the known binding constants should produce a model better suited for the absolute binding affinity predictions.

Acknowledgement

The authors thank Georgy Cheremovskiy from Moscow Institute of Physics and Technology for the initial development of the potential, and Georgy Derevyanko from Concordia University who proposed the initial formulation of the optimization problem. The authors also thank Valentin Gordeliy from IBS Grenoble, and Petr Popov from MIPT Moscow for fruitful discussions during this work. This work was partially supported by RSF research project 14-14-00995.

References

[1] Lingle Wang, BJ Berne, and Richard A Friesner. On achieving high accuracy and reliability in the calculation of relative protein–ligand binding affinities. *Proc. Natl. Acad. Sci. U. S. A.*, 109(6):1937–1942, 2012.

[2] Nadine Homeyer and Holger Gohlke. FEW: A workflow tool for free energy calculations of ligand binding. *J. Comput. Chem.*, 34(11):965–973, 2013.

[3] Lingle Wang, Yujie Wu, Yuqing Deng, Byungchan Kim, Levi Pierce, Goran Krilov, Dmitry Lupyan, Shaughnessy Robinson, Markus K. Dahlgren, Jeremy Greenwood, Donna L. Romero, Craig Masse, Jennifer L. Knight, Thomas Steinbrecher, Thijs Beuming, Wolfgang Damm, Ed Harder, Woody Sherman, Mark Brewer, Ron Wester, Mark Murcko, Leah Frye, Ramy Farid, Teng Lin, David L. Mobley, William L. Jorgensen, Bruce J. Berne, Richard A. Friesner, and Robert Abel. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.*, 137(7):2695–2703, 2015.

[4] Renxiao Wang, Yipin Lu, and Shaomeng Wang. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.*, 46(12):2287–2303, 2003.

[5] Douglas B Kitchen, Hélène Decornez, John R Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: Methods and applications. *Nat. Rev. Drug Discovery*, 3(11):935–949, 2004.

[6] Gregory L. Warren, C. Webster Andrews, Anna-Maria Capelli, Brian Clarke, Judith LaLonde, Millard H. Lambert, Mika Lindvall, Neysa Nevins, Simon F. Semus, Stefan Senger, Giovanna Tedesco, Ian D. Wall, James M. Woolven, Catherine E. Peishoff, , and Martha S. Head. A critical assessment of docking programs and scoring functions. *J. Med. Chem.*, 49(20):5912–5931, 2006.

[7] Campbell McInnes. Virtual screening strategies in drug discovery. *Curr. Opin. Chem. Biol.*, 11(5):494–502, 2007.

[8] Tiejun Cheng, Xun Li, Yan Li, Zhihai Liu, and Renxiao Wang. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.*, 49(4):1079–1093, 2009.

[9] R. D. Smith, Jr. Dunbar, J. B., P. M. Ung, E. X. Esposito, C. Y. Yang, S. Wang, and H. A. Carlson. Csar benchmark exercise of 2010: Combined evaluation across all submitted scoring functions. *J. Chem. Inf. Model.*, 51:2115–2131, 2011.

[10] Kelly L Damm-Ganamet, Richard D Smith, James B Dunbar Jr, Jeanne A Stuckey, and Heather A Carlson. Csar benchmark exercise 2011–2012: Evaluation of results from docking and relative ranking of blinded congeneric series. *J. Chem. Inf. Model.*, 53(8):1853–1870, 2013.

[11] Zheng Zheng and Kenneth M. Merz. Development of the knowledge-based and empirical combined scoring algorithm (kecsa) to score protein–ligand interactions. *J. Chem. Inf. Model.*, 53(5):1073–1083, 2013.

[12] Jie Liu and Renxiao Wang. Classification of current scoring functions. *J. Chem. Inf. Model.*, 55(3):475–482, 2015.

[13] Yan Li, Li Han, Zhihai Liu, and Renxiao Wang. Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *J. Chem. Inf. Model.*, 54(6):1717–1736, 2014.

[14] Jie Liu and Renxiao Wang. Classification of current scoring functions. *J. Chem. Inf. Model.*, 55(3):475–482, 2015.

[15] Bernard R Brooks, Robert E Bruccoleri, Barry D Olafson, David J States, S Swaminathan, and Martin Karplus. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.*, 4(2):187–217, 1983.

[16] William L. Jorgensen, David S. Maxwell, and Julian Tirado-Rives. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.*, 118(45):11225–11236, 1996.

[17] Todd JA Ewing, Shingo Makino, A Geoffrey Skillman, and Irwin D Kuntz. Dock 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.*, 15(5):411–428, 2001.

[18] David A Case, Thomas E Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J Woods. The amber biomolecular simulation programs. *J. Comput. Chem.*, 26(16):1668–1688, 2005.

[19] Berk Hess, Carsten Kutzner, David Van Der Spoel, and Erik Lindahl. Gromacs 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.*, 4(3):435–447, 2008.

[20] Bernd Kuhn, Paul Gerber, Tanja Schulz-Gasch, and Martin Stahl. Validation and use of the mm-pbsa approach for drug discovery. *J. Med. Chem.*, 48(12):4040–4048, 2005.

[21] Prasad Chaskar, Vincent Zoete, and Ute F. Röhrig. Toward on-the-fly quantum mechanical/molecular mechanical (qm/mm) docking: Development and benchmark of a scoring function. *J. Chem. Inf. Model.*, 54(11):3137–3152, 2014.

[22] Hans-Joachim Böhm. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comput.-Aided Mol. Des.*, 8(3):243–256, 1994.

[23] Matthew D Eldridge, Christopher W Murray, Timothy R Auton, Gaia V Paolini, and Roger P Mee. Empirical scoring functions: I. the development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.*, 11(5):425–445, 1997.

[24] Renxiao Wang, Luhua Lai, and Shaomeng Wang. Further development and validation of empirical scoring functions for structure-based binding affinity prediction. *J. Comput.-Aided Mol. Des.*, 16(1):11–26, 2002.

[25] Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis, and Peter S. Shenkin. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.*, 47(7):1739–1749, 2004.

[26] Oliver Korb, Thomas Stutzle, and Thomas E Exner. Empirical scoring functions for advanced protein–ligand docking with plants. *J. Chem. Inf. Model.*, 49(1):84–96, 2009.

[27] Yan Li, Zhihai Liu, Jie Li, Li Han, Jie Liu, Zhixiong Zhao, and Renxiao Wang. Comparative assessment of scoring functions on an updated benchmark: 1. compilation of the test set. *J. Chem. Inf. Model.*, 54(6):1700–1716, 2014.

[28] Oleg Trott and Arthur J Olson. Autodock vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, 31(2):455–461, 2010.

[29] Rodrigo Quiroga and Marcos A Villarreal. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PloS one*, 11(5):e0155183, 2016.

[30] Cheng Wang and Yingkai Zhang. Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *J. Comput. Chem.*, 38(3):169–177, 2017.

[31] Ingo Muegge and Yvonne C Martin. A general and fast scoring function for protein-ligand interactions: a simplified potential approach. *J. Med. Chem.*, 42(5):791–804, 1999.

[32] Wijnand Mooij and Marcel L Verdonk. General and targeted statistical potentials for protein–ligand interactions. *Proteins: Struct., Funct., Bioinf.*, 61(2):272–287, 2005.

[33] Sheng-You Huang and Xiaoqin Zou. Mean-force scoring functions for protein–ligand binding. *Annu. Rep. Comput. Chem.*, 6:280–296, 2010.

[34] Hongyi Zhou and Jeffrey Skolnick. Goap: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.*, 101(8):2043–2052, 2011.

[35] Sheng-You Huang and Xiaoqin Zou. Scoring and lessons learned with the csar benchmark using an improved iterative knowledge-based scoring function. *J. Chem. Inf. Model.*, 51(9):2097–2106, 2011.

[36] Gerd Neudert and Gerhard Klebe. Dsx: a knowledge-based scoring function for the assessment of protein–ligand complexes. *J. Chem. Inf. Model.*, 51(10):2731–2745, 2011.

[37] Petr Popov and Sergei Grudinin. Knowledge of native protein-protein interfaces is sufficient to construct predictive models for the selection of binding candidates. *J. Chem. Inf. Model.*, 55(10):2242–55, Oct 2015.

[38] Zhiqiang Yan and Jin Wang. Incorporating specificity into optimization: evaluation of spa using csar 2014 and casf 2013 benchmarks. *J. Comput.-Aided Mol. Des.*, 30(3):219–227, 2016.

[39] Holger Gohlke, Manfred Hendlich, and Gerhard Klebe. Knowledge-based scoring function to predict protein-ligand interactions. *J. Mol. Biol.*, 295(2):337–356, 2000.

[40] Sheng-You Huang and Xiaoqin Zou. Inclusion of solvation and entropy in the knowledge–based scoring function for protein–ligand interactions. *J. Chem. Inf. Model.*, 50(2):262–273, 2010.

[41] Christoph Sotriffer. Scoring functions for protein–ligand interactions. *Protein-Ligand Interactions, First Edition*, pages 237–263, 2012.

[42] Sarah L. Kinnings, Nina Liu, Peter J. Tonge, Richard M. Jackson, Lei Xie, and Philip E. Bourne. A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *J. Chem. Inf. Model.*, 51(2):408–419, 2011.

[43] David Zilian and Christoph A. Sotriffer. Sfcscorerf: A random forest-based scoring function for improved affinity prediction of protein–ligand complexes. *J. Chem. Inf. Model.*, 53(8):1923–1933, 2013.

[44] Guo-Bo Li, Ling-Ling Yang, Wen-Jing Wang, Lin-Li Li, and Sheng-Yong Yang. Id-score: A new empirical scoring function based on a comprehensive set of descriptors related to protein–ligand interactions. *J. Chem. Inf. Model.*, 53(3):592–600, 2013.

[45] Joffrey Gabel, Jérémy Desaphy, and Didier Rognan. Beware of machine learning-based scoring functions—on the danger of developing black boxes. *J. Chem. Inf. Model.*, 54(10):2807–2815, 2014.

[46] Christoph Sotriffer and Hans Matter. *Virtual Screening: Principles, Challenges, and Practical Guidelines*. Wiley Online Library, 10.1002/9783527633326.ch7 edition, 2011.

[47] Jason B Cross, David C Thompson, Brajesh K Rai, J Christian Baber, Kristi Yi Fan, Yongbo Hu, and Christine Humblet. Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *J. Chem. Inf. Model.*, 49(6):1455–1474, 2009.

[48] Jui-Hua Hsieh, Shuangye Yin, Shubin Liu, Alexander Sedykh, Nikolay V Dokholyan, and Alexander Tropsha. Combined application of cheminformatics and physical force field-based scoring functions improves binding affinity prediction for csar data sets. *J. Chem. Inf. Model.*, 51(9):2027–2035, 2011.

[49] Heather A Carlson, Richard D Smith, Kelly L Damm-Ganamet, Jeanne A Stuckey, Aqeel Ahmed, Maire A Convery, Donald O Somers, Michael Kranz, Patricia A Elkins, Guanglei Cui, Catherine E Peishoff, Millard H Lambert, and James B Dunbar, Jr. Csar 2014: A benchmark exercise using unpublished data from pharma. *J. Chem. Inf. Model.*, May 2016.

[50] Sergei Grudinin, Petr Popov, Emilie Neveu, and Georgy Cheremovskiy. Predicting binding poses and affinities in the csar 2013-2014 docking exercises using the knowledge-based convex-pl potential. *J. Chem. Inf. Model.*, 56(6):1053–1062, Nov 2016.

[51] Marc F. Lensink, Sameer Velankar, Andriy Kryshtafovych, Sheng-You Huang, D. Schneidman-Duhovny, Andrej Sali, J. Segura, N.s Fernandez-Fuentes, S. Viswanath, R. Elber, Sergei Grudinin, Petr Popov, Emilie Neveu, Hasup Lee, M. Baek, S. Park, L. Heo, G. Rie Lee, C. Seok, S. Qin, Hongyi Zhou, David W Ritchie, B. Maigret, Marie-Dominique Devignes, Anisah W. Ghoorah, Mieczyslaw Torchala, Raphael AG Chaleil, Paul A. Bates, E.t Ben-Zeev, M. Eisenstein, S.S. Negi, Zhiping Weng, Thom Vreven, Brian G. Pierce, T. M. Borrman, J. Yu, F. Ochsenbein, Raphael Guerois, A. Vangone, Joao PGLM Rodrigues, G. van Zundert, M. Nellen, L. Xue, E. Karaca, A.S.J. Melquiond, K. Visscher, Panagiotis L Kastritis, Alexandre M. J. J. Bonvin, X. Xu, L. Qiu, C. Yan, J. Li, Z. Ma, J. Cheng, X. Zou, Y. Shen, L.X. Peterson, H. Kim, A. Roy, X. Han, J. Esquivel-Rodriguez, D. Kihara, X. Yu, N.J. Bruce, J.C. Fuller, R.C. Wade, I. Anishchenko, P.J. Kundrotas, I.A. Vakser, K. Imai, K.i Yamada, T. Oda, T. T. Nakamura, K. Tomii, C. Pallara, M. Romero-Durana, B. Jimenez-Garcia, I.H. Moal, J. Fernandez-Recio, J. Young Joung, J. Yun Kim, K.g Joo, J. Lee, D. Kozakov, S. Vajda, S. Mottarella, D.R. Hall, D. Beglov, A. Mamonov, B. Xia, T. Bohnuud, C.A. Del Carpio, E. Ichiishi, N. Marze, D. Kuroda, S.S. Roy Burman, J.J. Gray, E. Chermak, L. Cavallo, R. Oliva, A. Tovchigrechko, and S.J. Wodak. Prediction of homo- and hetero-protein complexes by ab-initio and template-based docking: a CASP-CAPRI experiment. *Proteins: Struct., Funct., Bioinf.*, 2016.

[52] Sergei Grudinin, Maria Kadukova, Andreas Eisenbarth, Simon Marillet, and Fr'ed'eric Cazals. Predicting binding affinities for protein - ligand complexes in the 2015 d3r grand challenge using a physical model with a ridge regression parameter estimation. *J. Comput.-Aided Mol. Des.*

[53] Maria Kadukova and Sergei Grudinin. Docking of small molecules to farnesoid X receptors using AutoDock Vina with the Convex-PL potential : lessons learned from D3R Grand Challenge 2. *J. Comput.-Aided Mol. Des.*, 2017.

[54] San Diego: Dassault Systemes. *Ref. Dassault Systemes BIOVIA, Discovery Studio Modeling Environment, Release 2017*, 2016.

[55] André Krammer, Paul D Kirchhoff, X Jiang, CM Venkatachalam, and Marvin Waldman. Ligscore: a novel scoring function for predicting binding affinities. *J. Mol. Graphics Modell.*, 23(5):395–407, 2005.

[56] Daniel K Gehlhaar, Gennady M Verkhivker, Paul A Rejto, Christopher J Sherman, David R Fogel, Lawrence J Fogel, and Stephan T Freer. Molecular recognition of the inhibitor ag-1343 by hiv-1 protease: conformationally flexible docking by evolutionary programming. *Chem. Biol. (Oxford, U. K.)*, 2(5):317–324, 1995.

[57] Ajay N Jain. Scoring noncovalent protein-ligand interactions: a continuous differentiable function tuned to compute binding affinities. *J. Comput.-Aided Mol. Des.*, 10(5):427–440, 1996.

[58] Ingo Muegge. A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. In *Virtual Screening: An Alternative or Complement to High Throughput Screening?*, pages 99–114. Springer, 2000.

[59] Ingo Muegge. Effect of ligand volume correction on pmf scoring. *J. Comput. Chem.*, 22(4):418–425, 2001.

[60] Hans-Joachim Böhm. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from de novo design or 3d database search programs. *J. Comput.-Aided Mol. Des.*, 12(4):309–309, 1998.

[61] Christopher R Corbeil, Christopher I Williams, and Paul Labute. Variability in docking success rates due to dataset preparation. *J. Comput.-Aided Mol. Des.*, 26(6):775–786, 2012.

[62] Paul Labute. The generalized born/volume integral implicit solvent model: estimation of the free energy of hydration using london dispersion instead of atomic surface area. *J. Comput. Chem.*, 29(10):1693–1698, 2008.

[63] Junichi Goto, Ryoichi Kataoka, Hajime Muta, and Noriaki Hirayama. Asedock-docking based on alpha spheres and excluded volumes. *J. Chem. Inf. Model.*, 48(3):583–590, 2008.

[64] Sheng-You Huang and Xiaoqin Zou. An iterative knowledge-based scoring function for protein–protein recognition. *Proteins: Struct., Funct., Bioinf.*, 72(2):557–579, 2008.

[65] Gwo-Yu Chuang, Dima Kozakov, Ryan Brenke, Stephen R Comeau, and Sandor Vajda. Dars (decoys as the reference state) potentials for protein-protein docking. *Biophys. J.*, 95(9):4217–4227, 2008.

[66] Vladimir N Maiorov and Gordon M Crippen. Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.*, 227(3):876–888, 1992.

[67] Jian Qiu and Ron Elber. Atomically detailed potentials to recognize native and approximate protein structures. *Proteins: Struct., Funct., Bioinf.*, 61(1):44–55, 2005.

[68] Dror Tobi and Ivet Bahar. Optimal design of protein docking potentials: Efficiency and limitations. *Proteins: Struct., Funct., Bioinf.*, 62(4):970–981, 2006.

[69] Myong-Ho Chae, Florian Krull, Stephan Lorenzen, and Ernst-Walter Knapp. Predicting protein complex geometries with a neural network. *Proteins: Struct., Funct., Bioinf.*, 78(4):1026–1039, 2010.

[70] Emilie Neveu, David W Ritchie, Petr Popov, and Sergei Grudinin. Pepsi-dock: a detailed data-driven protein-protein interaction potential accelerated by polar fourier correlation. *Bioinformatics*, 32(17):i693–i701, Sep 2016.

[71] V. Vapnik. *The nature of statistical learning theory*. Springer, 2000.

[72] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, 2004.

[73] Ron Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model

selection. In *International Joint Conference on Artificial Intelligence*, volume 2, pages 1137–1145, 1995.

[74] S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge Univ Press, 2004.

[75] V. Vapnik. *Estimation of dependences based on empirical data*. Nauka, 1979.

[76] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop*, pages 276–285, 1997.

[77] J.C. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C.J.C. Burges, and A.J. Smola, editors, *Advances in Kernel Methods*. MIT press Cambridge, MA, 1998.

[78] Y.J. Lee and O.L. Mangasarian. RSVM: Reduced support vector machines. In *Proceedings of the First SIAM International Conference on Data Mining*, pages 00–07, 2001.

[79] Maria Kadukova and Sergei Grudinin. Knodle: A support vector machines-based automatic perception of organic molecules from 3d coordinates. *J. Chem. Inf. Model.*, 56(8):1410–1419, 2016.

[80] Gerd Neudert and Gerhard Klebe. fconv: Format conversion, manipulation and feature computation of molecular data. *Bioinformatics*, 27(7):1021–1022, 2011.

[81] Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The PDBbind Database: Methodologies And Updates. *J. Med. Chem.*, 48(12):4111–9, June 2005.

[82] Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *J. Med. Chem.*, 47(12):2977–80, June 2004.

[83] Petr Popov and Sergei Grudinin. Rapid determination of RMSDs corresponding to macromolecular rigid body motions. *J. Comput. Chem.*, 35(12):950–956, 2014.

[84] Symon Gathiaka, Shuai Liu, Michael Chiu, Huanwang Yang, Jeanne A Stuckey, You Na Kang, Jim Delproposto, Ginger Kubish, James B Dunbar, Heather A Carlson, et al. D3r grand challenge 2015: Evaluation of protein–ligand pose and affinity predictions. *J. Comput.-Aided Mol. Des.*, 30(9):651–668, 2016.

[85] Min-yi Shen and Andrej Sali. Statistical potential for assessment and prediction of protein structures. *Protein Sci.*, 15(11):2507–2524, 2006.

[86] Hongyi Zhou and Yaoqi Zhou. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.*, 11(11):2714–2726, 2002.

[87] Ram Samudrala and John Moult. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.*, 275(5):895–916, 1998.

[88] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. Blast+: architecture and applications. *BMC Bioinf.*, 10(1):421, 2009.

	Quality 1	Quality 2	Quality 3
	Native poses included		
Top 1	80.5	88.2	92.3
Top 2	86.2	91.8	94.9
Top 3	89.7	93.3	95.9
	Native poses excluded		
Top 1	71.3 (77.7)	86.2 (88.0)	91.3
Top 2	78.5 (85.5)	89.2 (91.1)	93.8
Top 3	83.6 (91.1)	91.8 (93.7)	94.9
	Poses with RMSD < 1 Å excluded		
Top 1	-	57.4 (58.6)	75.4
Top 2	-	70.1 (71.6)	84.1
Top 3	-	81.5 (83.2)	91.3

Table 1: CASF 2013 benchmark docking test success rates. The Quality columns 1–3 refer to the correctly predicted structures with RMSD < 1 Å, < 2 Å, and < 3 Å, respectively. Near-native poses within RMSD values of 1 Å and 2 Å were not available for 16 and 4 ligands, respectively. Therefore, we provide unnormalized success rates without parentheses and normalized success rates in parentheses.

	top 1			top 3			top 5		
Quality, Å	1	2	3	1	2	3	1	2	3
Success rate, %	58.3	69.7	76.5	70.8	78.8	88.2	75	84.8	88.2

Table 2: Success rates of finding top-1, top-3 and top-5 near-native poses within RMSD values of 1, 2 and 3 Å on a set of user-submitted decoys from the Stage 1 of D3R Grand Challenge 2.